Cohesion and Coherence Analyses of Extended Written Schizophrenic Discourse: An Exploratory Case Study

RICHARD WATSON TODD

School of Liberal Arts, King Mongkut's University of Technology Thonburi, Thailand
Author email: irictodd@kmutt.ac.th

Article information	Abstract
	Schizophrenic discourse is characterized by thought disorder, or a lack of coherence, prompting substantial research into identifying and measuring the incoherent discourse of schizophrenics. Much of this research has examined short extracts of elicited spoken data and used researcher judgments. This study examines the connectedness of naturally-occurring extensive written schizophrenic discourse using three methods of analysis: an automated analysis of cohesion, lexical cohesion analysis, and topic-based analysis focusing on propositional coherence. The results show that the schizophrenic texts are highly cohesive, especially in their use of connectives and the density of connections. The texts also show high proportions of topic shifts and a greater average distance of moves between concepts. These suggest that the topic structure of the schizophrenic texts consists predominantly of repetitive topic maintenance interspersed with short unrelated mini-topics, and it is this structure that manifests thought disorder.

INTRODUCTION

1. Analyzing coherence in schizophrenic discourse

Schizophrenia is manifested in impairments in language use (Caban et al., 2011). These involve so-called thought disorder, referring to a lack of coherence in discourse (Covington et al., 2005). Thus, a substantial proportion of research into schizophrenia has focused on identifying and measuring incoherent discourse.

The standard approach in this research is to compare the discourse produced by people diagnosed with schizophrenia with the discourse of 'normal' people, and this approach is taken in this paper. Following Allende-Cid et al. (2019), I will term the discourse and texts produced by people diagnosed with schizophrenia as schizophrenic discourse and schizophrenic texts. These terms are used purely for brevity and there is no intention of stigmatizing the authors of such discourse.

In their seminal review of research into schizophrenia, Rochester and Martin (1979) highlighted two key problematic assumptions underpinning the construct of thought disorder. First, it is

assumed that problems in cognitive processing can be inferred from problems in language production. Second, incoherence is assumed to be solely the fault of the speaker and has nothing to do with the listener. Research since the turn of the century has addressed these problems.

Much of the research into thought disorder in the twentieth century relied on researchers' judgments of incoherence, an approach still used in a few studies (e.g. Rogalski et al., 2010). Generally, however, more recent research has been based on one of three approaches that overcome the problems of judgments of thought disorder. First, some research has combined coherence analysis with neural imaging (e.g. Ditman & Kuperberg, 2007) to confirm relationships between language production and thought processing. Second, many studies have used theoretically-grounded analyses of specific aspects of coherence rather than broad judgments of overall coherence. For example, Meehan and MacLachlan (2010) investigated self construction through subject positioning in discourse as one specific aspect of coherence. Third, a common recent approach is to generate numerous metrics related to coherence, often through automated computational analyses, and to compare the metrics for schizophrenic discourse with those of 'normal' discourse (e.g. Allé et al., 2015; Just et al., 2019; Saavadra, 2010).

The most common findings from these studies concern reference and topic shifts. In schizophrenic discourse, the use of reference, especially pronouns, is found to be "abnormal" (Ditman & Kuperberg, 2010, p. 260), "ambiguous" (Iter et al., 2018, p. 137), or "atypical" (Badash, 2021, p. 141). Topic shifts in schizophrenic discourse are found to be far more frequent than in 'normal' discourse (Gernsbacher et al., 1999), "unexpected" (Badash, 2021, p. 141), or lacking signaling (Riou, 2015).

These recent approaches have been beneficial in producing rigorous justifiable findings that have pushed the field forward, but there are still some weaknesses in the research oeuvre into schizophrenic discourse. One potential problem is particularly apparent in the application of automated techniques to coherence. Coherence is usually distinguished from cohesion in that the former is implicit while the latter is explicit. This means that cohesion can be viewed as a property of texts (and is thus open to automated analyses), but coherence resides in people's interpretations of texts (Yule, 1996). Automated measures of coherence are therefore potentially problematic, and some studies using automated approaches which claim to be examining coherence are actually measuring cohesion (e.g. Just et al., 2019; Panicheva & Litvinova, 2019). One way around this is to ensure that any metrics used as measures of the connectedness of a text (an umbrella term covering cohesion and coherence) clearly focus on some specific aspect of connectedness. For instance, Allé et al. (2015) found that schizophrenic discourse manifested problems with thematic coherence to a much greater extent than problems with temporal coherence. Nevertheless, automated analyses claiming to measure coherence need to be treated with caution.

A second problem with the recent research into schizophrenic discourse is a tendency to focus on short spoken extracts. The vast majority of research into schizophrenic discourse has looked at spoken discourse, generally in the form of interviews or narratives, and in most cases the data has consisted of extracts of a few hundred words at most. While such research has been

productive, there is a risk that features associated with other forms of discourse may be overlooked, as is the case for written discourse where the findings can stand in contrast to those derived from spoken language (e.g., Panicheva & Litvinova (2019) found schizophrenic discourse to have fewer topic shifts in the second half of texts than in 'normal' discourse). There is then a need to redress the balance in data by examining longer extracts of written language.

One further possible issue is that nearly all recent research uses elicited data. Where the goal of the research is to aid diagnosis, this approach is appropriate since diagnostic procedures involve eliciting language from potential patients. However, the tendency to focus on elicited data means that little is known about schizophrenic discourse outside clinical contexts, again suggesting a need to redress the balance.

2. Purpose of the study

Given the issues with the previous research into schizophrenic discourse, this paper is an exploratory study focusing on the problems identified above. First, the data is naturally-occurring extended written discourse, the register probably least covered in previous studies. Second, three approaches to analyzing the data, each taking a different perspective on connectedness, will be used to shed light on the discourse features associated with schizophrenic discourse. By using such a methodology, this study aims to provide insights relevant to both the study of schizophrenia and discourse analysis. For schizophrenia, this research aims to identify differences in the discourse structure between schizophrenic texts and 'normal' texts which could be used as identifiers of thought disorder. For discourse analysis, the study aims to highlight the usefulness of the three different perspectives on connectedness.

METHODOLOGY

The primary data analyzed in this study is three academic articles written by a schizophrenic. The articles were analyzed in three ways representing different views on the nature of discourse:

1. An automated analysis focusing largely on cohesion producing numerous metrics.
2. A manual analysis focusing on lexical cohesion producing a map of the text.
3. A manual analysis focusing on propositional coherence producing several metrics and a map of the text.

In addition, three other academic articles by different authors but published in the same journals as the primary articles were also analyzed to provide a benchmark. Given the exploratory nature of the study and the dangers in drawing conclusions from a few texts, the benchmark analysis was conducted to allow interpretations of the nature of the primary texts to be drawn but was not used for definitive comparisons.

1. The data

Accessing naturally-occurring extensive written discourse produced by a person diagnosed with schizophrenia is problematic. There are collections of schizophrenic discourse (e.g. SAGE & Alexander Street Press, 2008) which could be used as data, but in these sources the written discourse is mostly retrospective accounts written by people who may no longer manifest the symptoms of schizophrenia. One source which appears to unequivocally represent schizophrenic discourse is a series of academic articles written by the same author and published in dubious open-access journals. These journals allow articles to be freely available for use under Creative Commons agreements. Nevertheless, given the potential risks, consent to use the articles and to give attribution to the author was sought and granted fulfilling ethical requirements.

Of the 23 articles in the series, the three articles chosen for this study are those where the author is open about his schizophrenia and about his refusal to take medication (e.g. "I did not use the drugs for schizophrenia (mostly Olanzapine) which were prescribed by the psychiatrist" Malekinejad, 2019, p. 478). The articles are argued opinion articles, rather than articles presenting research results and bibliographic details of the articles are given in Table 1. To illustrate the analyses, the article termed Addenda 4 is used as an example.

Although published in academic journals, the articles include some features not normally associated with academic writing:

- There are few headings in the article: Addenda 4, for instance, consists of: Introduction (covering the vast majority of the article) and Conclusion.
- There is extensive use of upper-case letters (e.g. "A LARGE PART OF DISCOVERED SCIENCE IS NOT REVEALED AND IS KEPT SECRET").
- Certain phrases are repeated frequently (e.g., the sentence in upper-case letters appears eight times in Addenda 4 with minor variations).
- Nearly all of the references which are not self-citations are to webpages such as Wikipedia pages and YouTube videos.
- Unusual highly personal information is included (e.g., "I do masturbation and I prefer it to any kind of sex").
- Self-references are very common (there are 93 instances of "I" in Addenda 4).

Overall, the articles generally combine arguments about the weaknesses of Iran (especially compared to the USA) with personal reflections.

Table 1
Overview of the articles representing schizophrenic discourse

Article reference	Short name	Length
Malekinejad, P. (2019)	Addenda 4	5,843
Malekinejad, P. (2020a)	Addenda 6	2,409
Malekinejad, P. (2020b)	Addenda 10	4,389

2. Data preparation

To focus on the structure of the extended writing in the articles, only the body of the text was used (in other words, the title, author details, abstract and references were not included in the analysis). Two of the three methods of analysis used (automated analysis of cohesion and lexical cohesion analysis) require the texts to be segmented into paragraphs. Although the articles are already presented in paragraphs, several of these are very long (e.g., one is over 2,000 words long) which might be a feature of schizophrenic written discourse. However, these long paragraphs could cause problems when looking at, for instance, the extent to which lexical items are repeated between paragraphs, an issue particularly relevant in lexical cohesion analysis.

Before conducting the analyses, these longer paragraphs were divided into shorter paragraphs each of 50-250 words based on connectives within the original paragraph. For example, in Addenda 4 "Also" is used twice at the start of paragraphs, but also appears six times in the middle of long paragraphs and so is used as a paragraph divider.

3. The benchmark articles

To enable interpretations of the results for the articles representing extended written schizophrenic discourse, three further articles from the same genre (i.e., academic articles published in the same journals or in journals where Malekinejad had published other articles in the series) and of similar length were also analyzed. Details of these benchmark articles are given in Table 2.

Table 2
Overview of the benchmark articles

Article reference	Short name	Length
Gogo, J. O. (2020)	Moral	5,722
Ndukauba, K. N. N. (2020)	French	2,949
Kware, A. A. (2020)	Poor	4,027

4. Data analysis

All articles were analyzed using three methods.

4.1 Automated analysis of cohesion

Automated analyses of schizophrenic discourse have become more common in recent years, perhaps because the tools underpinning such analyses are straightforward and quick to use, can process large amounts of data, and appear to avoid subjective interpretations. Since these tools process the surface features of the text input, they are far more likely to provide results concerning cohesion than coherence. Nevertheless, some tools which conduct fairly intricate analyses aim to provide coverage of several different aspects of cohesion which may link to expert ratings of the quality of discourse. One such tool is *Tool for the Automatic Analysis of Cohesion or TAACO* (see Kyle & Crossley, 2017). TAACO provides numerous indices which cover

four broad categories of cohesion: local cohesion such as connectives, text cohesion manifested through type-token ratio and givenness, cohesion of texts at a semantic level (for example, through latent Dirichlet allocation) which might be viewed as analyzing coherence (Crossley et al., 2019), and global cohesion through lexical and synonymy overlap between sentences and paragraphs (Crossley et al., 2016). For our purposes, the indices related to lexical diversity and sophistication are unlikely to be related to connectedness and so are not included in the analysis. Lexical diversity and sophistication concern the choice of wording to refer to concepts, whereas connectedness concerns the relationship between concepts irrespective of the choice of wording. The analysis therefore focuses on those indices measuring local, semantic and global cohesion (connectives, reference, and lexical and semantic overlap between sentences and paragraphs) which are most likely to be related to schizophrenic thought disorder.

4.2 Lexical cohesion analysis

Lexical cohesion is measured through 34 metrics in TAACO, but a different approach to lexical cohesion is to try to map the relationships between the various segments in the text. This is the approach used in Hoey's (1991) lexical cohesion analysis (LCA). This method is based on pairwise similarity between segments in the text on the assumption that the greater the number of shared words between segments, the more connected the segments (Sanfilippo, 1998). In Hoey's original work, since he was using short texts, the segments he analyzed were sentences, meaning that he counted the number of content words which occurred in both sentences in each possible pair of sentences in the text. For the articles analyzed in this study, the number of sentences (e.g. Addenda 4 has 243 sentences) means that basing the LCA on sentences is impractical. Other research (e.g. Phillips, 1989) has suggested that, for purposes of lexical cohesion, discourse can be viewed as fractal and thus longer segments can be used. In this study, the paragraphs will be used as the segments for analysis.

The use of paragraphs as segments, however, raises a different issue. If every content word is counted in every paragraph, the number of links between segments will be very high. Rather than using every content word, in this study I am focusing on repetitions of keywords. Keywords are words whose frequency is noticeably higher in the text under study than in a comparative corpus which is taken as being indicative of their importance to the text (Scott & Tribble, 2006). In this study, keywords were identified by comparing the articles with the British National Corpus using *KeyBNC* (Graham, 2014). The words were ranked by log likelihood and the top content words were taken as keywords. Identifying cutoff points for words to be considered keywords is problematic. Using log likelihood values or associated probability values is unreliable given that these are heavily influenced by the size of the corpora (Pojanapunya & Watson Todd, 2018). Instead, since the goal of the research is to identify relationships between concepts, a more pragmatic approach was taken. Cutoff points for words to be considered keywords were set where the ranked list of words started to include several function words (rather than only content words) meaning that the keywords are restricted to content words.. The resulting keywords were further reduced by counting multiple keywords only appearing in fixed phrases (e.g. *depressive realism*) as single words. As a result, the number of keywords identified ranged from 18 to 29 and these were counted for repetition between paragraphs. In addition, the number of repetitions between paragraphs of types, rather than tokens, was counted given

the circularity of some of the text. In other words, if a keyword appears four times in paragraph 1 and three times in paragraph 2, this is counted as a single repetition (of the keyword as type) rather than each individual occurrence being counted separately.

The basic procedure, then, is to count the number of keywords which are repeated between each possible pair of paragraphs, with each repetition being termed a link. Doing this produces a table giving the number of links between each of the possible pairs of paragraphs. The next stage is to set a threshold for the number of links above which the two paragraphs are considered bonded. This threshold should be set at a level that identifies 2-10% of possible pairs of segments as being bonded (Watson Todd, 2016). In the case of Addenda 4, setting the threshold at 4 links gives 3.7% of the pairs as bonded. Finally, a map of the text can be drawn showing these bonds to represent how the whole text is connected.

4.3 Topic-based analysis

Topic-based analysis (TBA) focuses on the propositional or ideational coherence (Redeker, 1990) of a text by creating a schematic structure of the concepts in the text and tracing the sequence in which they appear in the discourse (Watson Todd, 2003; 2016). This results in a map of the text and allows metrics for coherence to be generated.

The first stage in TBA is to identify the main concepts in the text. The same concepts identified through a keyword analysis for LCA were used. The next stage is to create a hierarchical schematic structure linking these concepts. This structure is text-specific and uses both generic semantic relations between concepts and text-specific non-classical relations which are created by the discourse. Because of this, there is an element of subjectivity in creating the schematic structure. However, by following the procedures in creating the schema illustrated in Figure 1, the extent to which the element of subjectivity is likely to influence the results can be reduced so that its impact on reliability is minimal.

The process of creating the hierarchy for a text can be illustrated by looking at four of the concepts from Addenda 4: *donkey, nuclear, science* and *secret*. From a generic semantic perspective *nuclear* and *science* are related especially in the phrase *nuclear science* (which occurs 71 times in the Corpus of Contemporary American English with an MI score of 2.56) implying that *nuclear* is subordinate to *science*. In general English use, there is a possible loose relation between *nuclear* and *secret*, but this is less clear. In the text, however, at several points *science* and *secret* are clearly linked (as in the quotation in capital letters above) with *secret* represented as a property of (and thus subordinate to) *science*. Finally, *donkey* has no generic or text-specific relations to the other three concepts. From these, we can create a map showing these relations as in Figure 1.
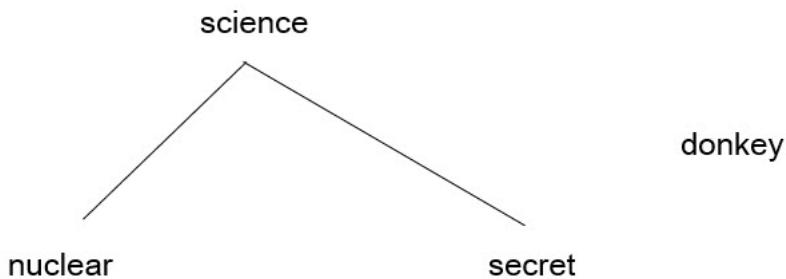
Figure 1 Map showing relations between four concepts

Once a full schematic map of the concepts in the text have been drawn up, the sequence in which they appear in the discourse can be added to the map. The goal here is to see how the discourse progresses through the schematic semantic space of the hierarchy. For example, three of the concepts in Figure 1 appear in the same sentence in the text ("As another example of this part of science which is not revealed to public and is kept secret, I can mention the construction of nuclear power plants."). These would be represented as moves through the map as in Figure 2 where the numbers show the sequence of moves.
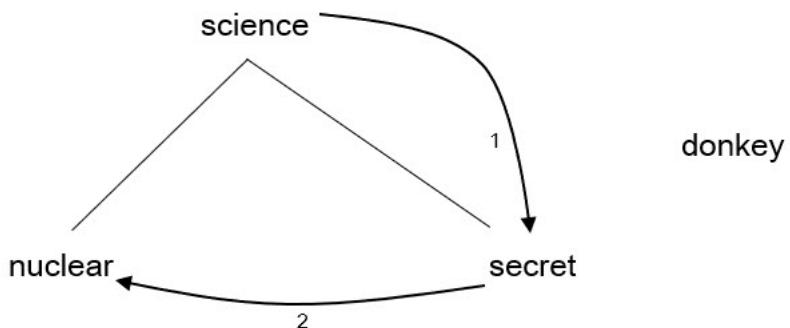

Figure 2 Map showing relations between four concepts with moves added

From the map, certain metrics can be generated. Distances between concepts in the schematic hierarchy can be given values so that, for example, a move between a superordinate and a direct subordinate would be given a value of 1. Moves between co-subordinates are also assigned a value of 1. Moves between completely unrelated concepts where there is no path between them (representing coherence breaks) would be given a value greater than the maximum value for any move between two sequential linked concepts. The frequency at which concepts and moves are recycled can also be counted.

RESULTS

1. Automated cohesion analysis

The metrics produced by TAACO for the categories of connectives, lexical overlap and semantic overlap for the schizophrenic texts and the benchmark texts are given in Table 3. The metrics are sequenced based on the percentage difference of the values between the two sets of texts with those at the top being the metrics where the schizophrenic texts most noticeably scored higher than the benchmark texts on average. In addition, each metric is categorized as falling within one of four broad categories: connectives, reference (such as use of demonstratives), sentence (such as lexical reiteration across sentences), and paragraph (such as lexical reiteration across paragraphs).

Four main patterns emerge from Table 3. First, many of the differences between the two sets of texts in terms of cohesion are negligible – for over half of the metrics, including most concerning paragraph overlap, the difference between the two average values is less than 20% - suggesting that cohesion is not a good indicator of the thought disorder of schizophrenics.

Second, for those points where there are some differences, the schizophrenic texts use a relatively high proportion of connectives (with the exception of negative connectives e.g. *however, alternatively*). 6 of the 13 metrics where the schizophrenic texts score at least 20% higher than the benchmark texts concern connectives. Connectives are perhaps the most noticeable form of cohesion and there is some evidence that they may generally be overused (Jones, 2010) since authors may believe that using, say, positive logical connectives makes their writing more convincing. The relative dearth of negative connectives in the schizophrenic texts suggests that the arguments may be largely uni-directional.

Third, the use of demonstratives, either as pronouns or as determiners followed by noun phrases, is far more frequent in the schizophrenic texts. Previous research into the use of determiners in academic discourse has shown that they are most often used to refer to antecedent complete clauses (Gray, 2010). This usage is also apparent in the schizophrenic texts where common uses of *this* are in the phrases *this is because, this fact* and *in this way* with the function of providing the author's interpretations of his previous statements. A substantial proportion of the schizophrenic texts consists of such elaborations within the framework of a single topic.

Fourth, for sentence overlap the schizophrenic texts have frequent repetition of nouns in succeeding sentences, but no such repetition for verbs and little evidence of broader similarities between succeeding sentences (as measured by, say, latent semantic analysis). This suggests extensive repetitive topic maintenance across sentences.

Table 3
TAACO metrics of cohesion for the schizophrenic and benchmark texts

TAACO metric	Add 4	Add 6	Add 10	French	Moral	Poor	% difference between schizophrenic and benchmark texts	Focus of metric
Number of demonstratives followed by a noun phrase	0.014	0.014	0.016	0.004	0.005	0.002	120.0	Reference
Number of order words	0.006	0.002	0.004	0.003	0.001	0.001	82.4	Connectives
Number of coordinating conjuncts	0.007	0.006	0.004	0.003	0.003	0.002	72.0	Connectives
Number of temporal connectives	0.014	0.012	0.010	0.006	0.006	0.008	57.1	Connectives
Number of noun lemma types that occur at least once in the next sentence (sentence normed)	1.083	1.000	0.980	0.503	0.791	0.493	52.6	Sentence
Average sentence to sentence overlap of noun synonyms	1.464	1.154	1.201	0.588	1.081	0.636	49.4	Sentence
Number of demonstratives	0.032	0.027	0.039	0.018	0.021	0.022	46.5	Reference
Number of positive causal connectives	0.027	0.023	0.030	0.019	0.019	0.017	37.0	Connectives
Number of content lemma types that occur at least once in the next sentence (sentence normed)	1.732	1.470	1.779	0.933	1.481	1.101	34.5	Sentence
Adjacent paragraph overlap verb lemmas	0.256	0.181	0.262	0.209	0.142	0.156	31.8	Paragraph
Number of sentences with ANY noun lemma overlap with next sentence	0.632	0.581	0.533	0.412	0.515	0.346	31.3	Sentence
Number of positive logical connectives	0.020	0.017	0.019	0.010	0.014	0.017	30.9	Connectives
Number of lexical items functioning as subordinators	0.019	0.017	0.026	0.015	0.015	0.016	29.6	Connectives
Number of sentences with ANY content lemma overlap with next sentence	0.728	0.641	0.698	0.576	0.728	0.544	11.2	Sentence
Number of demonstratives functioning as a noun phrase	0.018	0.013	0.023	0.014	0.015	0.020	9.7	Reference
Number of positive connectives	0.078	0.076	0.086	0.069	0.075	0.074	9.6	Connectives
Number of sentence linking words	0.023	0.022	0.023	0.014	0.020	0.028	9.2	Connectives

TAACO metric	Add 4	Add 6	Add 10	French	Moral	Poor	% difference between schizophrenic and benchmark texts	Focus of metric
Number of reason and purpose words	0.017	0.017	0.014	0.010	0.014	0.020	8.7	Connectives
Number of addition words	0.045	0.043	0.053	0.043	0.046	0.042	7.4	Connectives
Number of causal connectives	0.011	0.008	0.012	0.010	0.009	0.010	6.7	Connectives
Average latent semantic analysis cosine similarity between all adjacent sentences	0.398	0.308	0.325	0.288	0.412	0.282	4.9	Sentence
Number of sentences with ANY verb lemma overlap with next sentence	0.234	0.222	0.312	0.152	0.294	0.300	2.9	Sentence
Number of conjunctions	0.041	0.039	0.047	0.041	0.043	0.040	2.4	Connectives
Number of logical connectives	0.035	0.027	0.035	0.023	0.033	0.039	2.1	Connectives
Adjacent paragraph overlap verb lemmas (paragraph normed)	2.838	1.813	2.875	3.222	1.720	2.458	1.7	Paragraph
Average latent Dirichlet allocation divergence score between all adjacent paragraphs	0.993	0.997	0.992	0.977	0.990	0.988	0.9	Paragraph
Number of all connectives	0.080	0.072	0.084	0.067	0.081	0.086	0.9	Connectives
Average latent Dirichlet allocation divergence score between all adjacent paragraphs (with a two-paragraph span)	0.997	0.998	0.996	0.984	0.995	0.993	0.6	Paragraph
Number of verb lemma types that occur at least once in the next sentence (sentence normed)	0.272	0.231	0.382	0.170	0.336	0.378	0.1	Sentence
Number of positive intentional connectives	0.003	0.002	0.008	0.005	0.005	0.003	0.0	Connectives
Binary adjacent paragraph overlap content lemmas	1.000	1.000	1.000	1.000	1.000	1.000	0.0	Paragraph
Average latent Dirichlet allocation divergence score between all adjacent sentences	0.939	0.957	0.948	0.951	0.966	0.930	-0.1	Sentence

TAACO metric	Add 4	Add 6	Add 10	French	Moral	Poor	% difference between schizophrenic and benchmark texts	Focus of metric
Average sentence to sentence overlap of verb synonyms	0.410	0.359	0.653	0.267	0.638	0.521	-0.3	Sentence
Average latent Dirichlet allocation divergence score between all adjacent sentences (with a two-sentence span)	0.969	0.973	0.971	0.973	0.982	0.975	-0.6	Sentence
Average latent semantic analysis cosine similarity between all adjacent sentences (with a two-sentence span)	0.744	0.670	0.708	0.706	0.762	0.679	-1.2	Sentence
Average paragraph to paragraph overlap of noun synonyms	12.216	8.668	10.406	12.056	8.920	10.958	-2.0	Paragraph
Average word2vec similarity score between all adjacent sentences (with a two-sentence span)	0.909	0.891	0.903	0.919	0.937	0.904	-2.1	Sentence
Binary adjacent paragraph overlap verb lemmas	0.919	0.938	1.000	1.000	0.920	1.000	-2.2	Paragraph
Average word2vec similarity score between all adjacent sentences	0.871	0.848	0.861	0.881	0.906	0.860	-2.6	Sentence
Average word2vec similarity score between all adjacent paragraphs (with a two-paragraph span)	0.934	0.914	0.935	0.964	0.974	0.965	-4.2	Paragraph
Adjacent paragraph overlap noun lemmas (paragraph normed)	5.216	3.750	4.375	5.056	3.840	5.042	-4.4	Paragraph
Average word2vec similarity score between all adjacent paragraphs	0.911	0.882	0.914	0.952	0.961	0.952	-5.7	Paragraph
Number of basic connectives	0.045	0.041	0.050	0.047	0.049	0.049	-6.4	Connectives
Number of additive connectives	0.056	0.047	0.056	0.049	0.059	0.062	-6.7	Connectives
Binary adjacent paragraph overlap noun lemmas	0.946	0.875	0.906	0.944	1.000	1.000	-7.7	Paragraph

TAACO metric	Add 4	Add 6	Add 10	French	Moral	Poor	% difference between schizophrenic and benchmark texts	Focus of metric
Adjacent paragraph overlap content lemmas (paragraph normed)	10.892	7.563	9.500	11.944	7.720	10.750	-8.4	Paragraph
Number of determiners	0.091	0.087	0.091	0.090	0.103	0.100	-8.5	Reference
Average latent semantic analysis cosine similarity between all adjacent paragraphs (with a two-paragraph span)	0.788	0.670	0.777	0.820	0.853	0.829	-11.3	Paragraph
Average paragraph to paragraph overlap of verb synonyms	8.514	7.125	11.813	12.000	8.140	12.708	-17.9	Paragraph
Average latent semantic analysis cosine similarity between all adjacent paragraphs	0.512	0.347	0.412	0.607	0.592	0.534	-30.8	Paragraph
Number of negative logical connectives	0.005	0.001	0.004	0.002	0.006	0.008	-46.2	Connectives
Number of opposition words	0.004	0.001	0.003	0.002	0.006	0.007	-60.9	Connectives
Number of negative connectives	0.007	0.002	0.005	0.006	0.011	0.017	-83.3	Connectives
Number of disjunctions	0.002	0.001	0.001	0.005	0.005	0.008	-127.3	Connectives

2. Lexical cohesion analysis

LCA focuses on similar issues to some of the metrics in TAACO, such as 'Adjacent paragraph overlap content lemmas', but, instead of calculating metrics, produces a map of the text. Counting the reiterations of key concepts between each possible pair of paragraphs, for the schizophrenic text Addenda 4 we find the bonded paragraphs given in Table 4.

Table 4
Bonds between pairs of paragraphs for LCA in Addenda 4

Bonds to earlier paragraphs		*Bonds to later paragraphs*
	1	
	2	28. 30, 32, 34, 37, 38
	3	
	4	
	5	6, 17, 32
5	6	17, 32
	7	
	8	32, 34
	9	
	10	11, 12, 22, 24, 25, 26, 32, 34,
10	11	12, 15, 22, 23, 24, 25, 26
10, 11	12	24, 25, 26
	13	14, 26,
13	14	22, 24, 26
11	15	22
	16	
5, 6	17	
	18	
	19	
	20	
	21	
10, 11, 14, 15	22	24, 25, 26, 37
11	23	
10, 11, 12, 14, 22	24	25, 26
10, 11, 12, 22, 24	25	26
10, 11, 12, 13, 14, 22, 24, 25	26	
	27	
2	28	30, 32, 34, 37, 38
	29	
2, 28	30	37
	31	
2, 5, 6, 8, 10, 28	32	34, 35, 37
	33	
2, 8, 10, 28, 32	34	37
32	35	
	36	
2, 22, 28, 30, 32, 34	37	38
2, 28, 37	38	
	39	

Table 4 shows that many of the paragraphs are fairly heavily linked to numerous other paragraphs suggesting high levels of lexical cohesion. There are a few paragraphs which are unbonded and these focus on specific topics tangentially related to the main argument. For example, paragraph 27 concerns the author's ideas about 'utopia' (identified as a key concept because of its frequency in the text but only appearing in this paragraph). The bonded paragraphs can be used to create the map of the text shown in Figure 3.

The LCA map in Figure 3 suggests a highly connected text. Most paragraphs have several bonds with other paragraphs, although there are a few paragraphs which are not included in the map since they are not bonded with any other paragraph. The bonds show a mix of close (such as between paragraphs 10, 11 and 12) and long-distance connections (such as between paragraphs 10 and 11 on the one hand and 25 and 26 on the other). Such apparent high levels of connectivity, however, are symptomatic of a text deviating from the norm. Most LCA text maps have several features that are not apparent in the LCA map for the schizophrenic text, namely:

- some highly connected local clusters of text segments among a sparse network of more distant connections;
- few long-distance bonds with most (for written texts) linking the start and end of the text;
- a clumpy overall pattern in the map (where there are several heavily connected local clusters separated from others).

These features of 'normal' texts can be seen in Figure 4 which shows the LCA map for the 'normal' text Poor. In this map, there are three local clusters (paragraphs 1-3; 8-16; and 18-25) with the first and last of these also connected heavily.
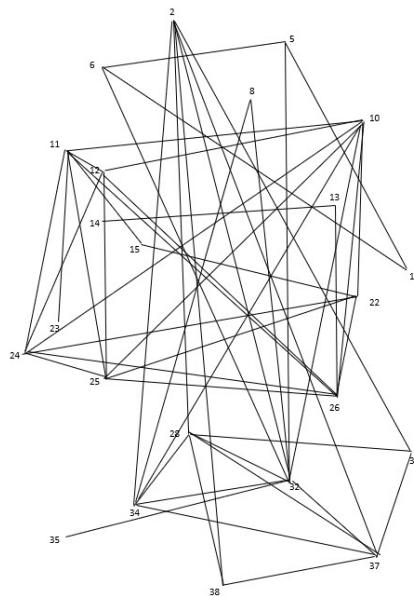

Figure 3 LCA map for Addenda 4
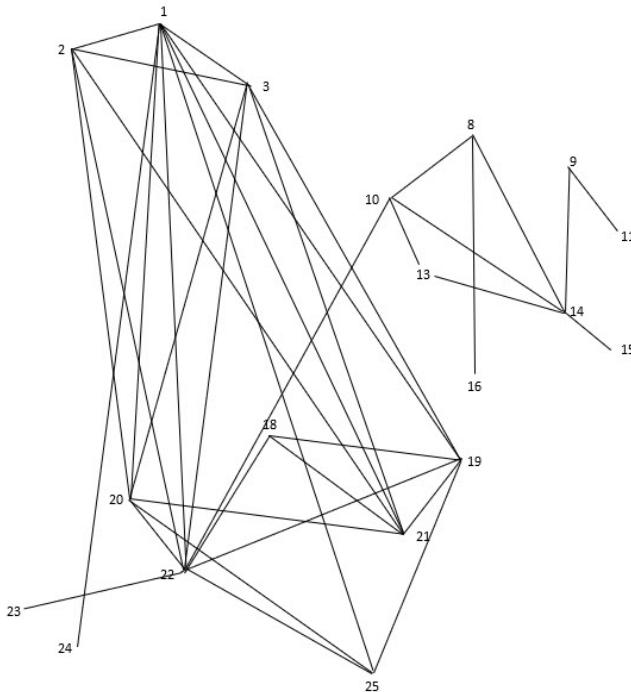
Figure 4 LCA map for Poor

These 'normal' features illustrate topic development through the discourse. The local clusters often indicate where a sub-topic is discussed in depth before moving on to the next sub-topic in the next local cluster. The few long-distance bonds often highlight the superordinate topic occurring in the introduction and conclusion.

The 'deviant' map in Figure 3 suggests that schizophrenic discourse is organized differently. Rather than moving through sub-topics, the schizophrenic discourse reiterates the same set of concepts at numerous points in the text creating a relatively high frequency of long-distance bonds. One potential indicator of schizophrenic discourse, then, may be a preference for topic maintenance (Crow, 1983) or topic recycling (Gardner, 1987) over topic drift (Hobbs, 1990) and topic progression (Dolón & Sánchez, 1999).

One further difference between the schizophrenic and 'normal' texts concerns those unbonded paragraphs which do not appear in the LCA maps. While the proportion of paragraphs which are unbonded is similar in the two sets of texts, all of the unbonded paragraphs in the 'normal' texts are linked to other paragraphs (albeit at a level not sufficient to be considered bonded). In the schizophrenic texts, on the other hand, several of the unbonded paragraphs are also unlinked, in other words, they contain no key concepts in common with any of the other paragraphs.

3. Topic-based analysis

Using the concepts from the keyword analysis, the schematic hierarchy shown in Figure 5 was constructed for Addenda 4. The hierarchy is somewhat subjective (although by following clear procedures the impact of this subjectivity is kept to a minimum) but has some interesting features. Most noticeably, there are three concepts (*schizophrenia, utopia,* and *donkey*) which are not related to any other concept in the hierarchy and another two concepts (*depressive realism* and *happy*) which are also isolated. The hierarchies for the other schizophrenic texts also include isolated concepts. For 'normal' texts, completely unrelated concepts are highly unusual and there are no isolated concepts in the TBA hierarchies for the three 'normal' texts. To see why this pattern has emerged, concordance plots for some of the concepts showing their location within Addenda 4 were constructed using *AntConc* 3.5.8 (Anthony, 2019). These are shown in Table 5.

Table 5
Concordance plots for selected concepts in the schizophrenic text

The concordance plots show that the concepts within the main structure of the hierarchy either appear scattered throughout the text (*science*) or cluster at several points in the text (*universities*). The completely unrelated concepts are highly localized in the text each appearing in a single paragraph (*utopia*). Finally, *depressive realism* and *happy* (which co-occur) are scattered but each occurrence is the only occurrence in that paragraph (unlike *universities*) suggesting that these concepts are never clearly developed.

Having created a hierarchy, the sequence in which the concepts appear in the text can be mapped onto this (following the guidelines in Author, 2016). The first 10 moves between concepts in Addenda 4 are shown in Figure 6. We can see that the initial moves are all within related concepts in the hierarchy (*USA* to the immediate hyponym of *politics* to the immediate superordinate of *Iran* to the hyponym of *nuclear*). Moves 6 and 7, however, represent clear shifts between unrelated concepts followed by a longer-distance move from *Voice of America* to *Iran*. Tracing such moves allows us to generate various metrics representing different aspects of coherence and these are shown in Table 6 for all texts.
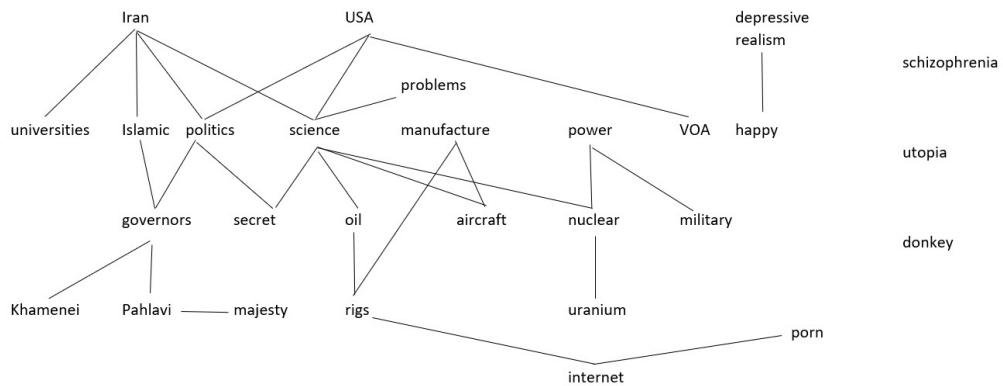
Figure 5 Schematic hierarchy of concepts in Addenda 4
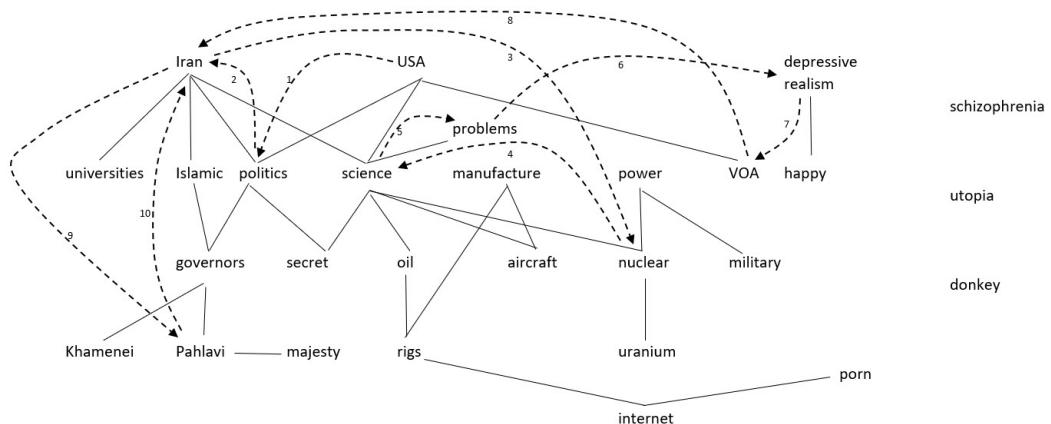
Figure 6 The first 10 moves between concepts in the schematic hierarchy for Addenda 4

Table 6

TBA metrics for the schizophrenic and 'normal' texts

Metric	Add 4	Add 6	Add 10	Poor	French	Moral
Average distance of moves	2.6	2.7	2.3	1.7	1.6	1.9
Proportion of moves as topic shifts	0.32	0.14	0.07	0	0	0
Number of moves per concept	3.2	3.7	4.1	5.2	8.5	8.4
Proportion of recycled concepts	0.74	0.89	0.85	1.00	1.00	0.95
Proportion of recycled moves	0.23	0.21	0.33	0.23	0.40	0.34

From Table 6, it can be seen that conducting a TBA of the schizophrenic text generates values for distance of moves and proportion of topic shifts markedly higher than in the 'normal' texts. These metrics are related as topic shift moves (such as from *utopia* to *USA*) have high distance values raising the average.

It is also noticeable that, based on the proportions of recycled concepts, nearly all of the concepts in the 'normal' texts appear at several points in the texts meaning that they are recycled. In contrast, in the schizophrenic texts, there are several concepts that appear at only one point in the text (they may be mentioned several times in one paragraph but do not appear elsewhere in the text as we saw with *utopia* in Addenda 4). These are instances of topic insert (Garcia & Joanette, 1997) where a short stretch of discourse on an unrelated topic is inserted into the main discourse.

Finally, the proportions of moves per concept are lower in the schizophrenic texts than the 'normal' texts. To some extent, this is a methodological artifact of following the guidelines of Watson Todd (2016). To avoid mapping an excessive number of moves onto the hierarchy, "only moves which represent progression in the discourse are mapped" (pp. 130-131). For parts of the schizophrenic texts, the discourse consists of a frequent circular toing and froing between two or three concepts where only the first move is counted. The low proportions of moves per concept in the schizophrenic texts are due to long stretches of topic maintenance. From the TBA, the key coherence features of schizophrenic texts are long stretches of topic maintenance interspersed with inserts on unrelated topics.

DISCUSSION

This study has investigated evidence for the impairments of language use associated with schizophrenia in naturally-occurring written academic text. Specifically, the lack of connectedness termed thought disorder has been analyzed using three different approaches: an automated analysis of cohesion producing metrics, a manual analysis of lexical cohesion producing a map of the text, and a manual analysis of propositional coherence producing both metrics and a map.

The two approaches focusing on cohesion both find extensive use of cohesion in the schizophrenic written text. The cohesion metrics for schizophrenic discourse are generally similar to 'normal' discourse, and the text map shows a single dense map covering most of the text. At first glance, these findings run counter to what we should expect for schizophrenic discourse manifesting thought disorder, and thus warrant deeper examination.

The TAACO metrics highlight the frequent use of a small set of positive connectives (especially *therefore* and *also*) and demonstratives. The use of these features implies consistency of argumentation suggesting that the schizophrenic texts exhibit topic maintenance as the dominant type of topic development. This emphasis on topic maintenance contrasts with the preferred pattern of steady topic drift or progression through sub-topics which is often found in model writing (Hobbs, 1990; Stede, 2012).

The single dense cluster of bonds in the LCA map for Figure 3 also suggests predominant topic maintenance, and also contrasts with the pattern found in many 'normal' texts of several clusters through the text, each representing a sub-topic, with a few long-distance bonds between clusters. However, there are several paragraphs that do not appear on the LCA map.

These focus on concepts that are unrelated to the main topic, for example, the paragraph concerning the author's ideas about *utopia*. The somewhat repetitive topic maintenance, then, is interrupted at several points by paragraphs on unrelated topics or topic insert. This pattern is similar to that found in Noël-Jorand et al.'s (1997) analysis of schizophrenic speech samples where the main discourse was "interspersed with unexpected 'language satellites' consisting of a secondary short and specific discourse which was also well planned but had no relevance to the main discourse" (p. 183).

As noted above, the differences in most of the TAACO metrics between the schizophrenic texts and the 'normal' texts are small. If the schizophrenic texts consisted solely of topic maintenance while the 'normal' texts consisted of topic drift, we might expect greater differences in the metrics. However, if the schizophrenic texts' topic maintenance were occasionally interrupted by coherence breaks to short unrelated topics as suggested by the LCA findings, the high levels of cohesion associated with topic maintenance would be reduced on average by the occasional coherence break, meaning that the values for the metrics based on averages would be closer to the values produced for a text which is predominantly topic drift. In this case, we would expect the TAACO metrics based on averages for the two sets of texts to be similar. In such a case, metrics based on amount of variation, such as standard deviation, would differ markedly for the two sets of texts. The 'normal' texts would have consistently moderate cohesion values throughout the texts showing little variation; the schizophrenic texts would have mostly high cohesion values with occasional very low values leading to a similar average but far greater variation. In such cases an automated cohesion analysis producing variation metrics would be better suited to identifying schizophrenic discourse.

The patterns identified from the cohesion analyses are perhaps more apparent in the coherence-focused TBA. In constructing the schematic hierarchy, several concepts (such as *utopia*) were disconnected from or loosely connected to the main hierarchical structure. The paragraphs concerning these disconnected topics are the unexpected language satellites of Noël-Jorand et al. (1997). Moves between the main structure and these disconnected concepts are responsible for both the high number of topic shifts and the high average distance of move. On the other hand, the lower values for moves per concept are an artifact of analyzing texts which predominantly consist of topic maintenance.

This study is severely limited in analyzing the discourse of a single person diagnosed with schizophrenia and it is not clear the extent to which the findings would apply to other authors diagnosed with schizophrenia. The difficulties in finding and collecting examples of authentic discourse written by people diagnosed with schizophrenia, especially unmedicated patients, mean that increasing the sample size beyond a single author is impossible. Acknowledging this constraint, there are two potentially important implications. First, the most basic interpretation of thought disorder, namely, that cohesion and coherence will be sparse in schizophrenic discourse, is not apparent from the findings. Although the topical structure of the schizophrenic texts deviates from 'normal' discourse, there is substantial evidence of connectedness in the texts, especially for cohesion. Such evidence of cohesion in a text, however, provides little information about the quality of the text (cf. Crossley & McNamara, 2010). In this case, it appears to be the prevalent patterns of topic development, namely, a predominant structure

of repetitive topic maintenance interspersed with several unrelated mini-topics, that characterizes thought disorder.

Second, it is important to supplement the basic results of the analyses with in-depth examination of the text. This is particularly the case for automated analyses which for reasons of practicality are understandably growing in popularity. It is only when we compare these results with the text itself and try to understand the reasons for why the results came out as they did that we start to understand the nature of thought disorder in schizophrenic discourse. The automated analysis produces results that suggest that frequent use of positive connectives and demonstratives and infrequent use of negative connectives characterize schizophrenic texts. While potentially useful, these findings highlight the prevalence of topic maintenance throughout the schizophrenic texts and overlook the key feature of interspersed unrelated language satellites. It is only when we conduct the manual LCA and TBA that this potential identifier of schizophrenic discourse is revealed. In this way, this study highlights the importance of supplementing quantitatively-oriented analyses with in-depth interpretations of discourse.

THE AUTHOR

Richard Watson Todd is an associate professor at King Mongkut's University of Technology Thonburi, Thailand. He has a PhD from the University of Liverpool, and is the author of numerous articles and several books, including Discourse Topics (John Benjamins, 2016). His research interests include text linguistics, corpus linguistics and educational innovations.
irictodd@kmutt.ac.th

REFERENCES

Allé, M. C., Potheegadoo, J., Köber, C., Schneider, P., Coutelle, R., Habermas, T., Danion, J., & Berna, F. (2015). Impaired coherence of life narratives of patients with schizophrenia. *Scientific Reports, 5*, Article 12934. https://doi.org/10.1038/srep12934

Allende-Cid, H., Zamora, J., Alfaro-Faccio, P., & Alonso-Sánchez, M. F. (2019). A machine learning approach for the automatic classification of schizophrenic discourse. *IEEE Access, 7*, 45544- 45553. https://doi.org/10.1109/ACCESS.2019.2908620

Anthony, L. (2019). *AntConc 3.5.8* (Windows). [Computer program]

Badash, M. (2021). Topic shifts in spontaneous interaction of speakers with schizophrenia: Cohesion and thematic structures. *Text & Talk, 41*(2), 141–163. https://doi.org/10.1515/text-2019-0113

Caban, A., Valle-Lisboa, J. C., Elvevåg, B., & Mizraji, E. (2011). Detecting order–disorder transitions in discourse: Implications for schizophrenia. *Schizophrenia Research, 131*(1-3), 157-164. https://doi.org/10.1016/j.schres.2011.04.026

Covington, M. A. He, C. Brown, C., Naçi, L., McClain, J. T., Fjordbak, B. S., Semple, J., & Brown, J. (2005). Schizophrenia and the structure of language: The linguist's view. *Schizophrenia Research, 77*(1), 85-98. https://doi.org/10.1016/j.schres.2005.01.016

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods, 51*, 14-27. https://doi.org/10.3758/

s13428-018-1142-4

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods, 48*(4), 1227-1237. https://doi.org/10.3758/s13428-015-0651-7

Crossley, S., & McNamara, D. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. *Proceedings of the Annual Meeting of the Cognitive Science Society, 32*.

Crow, B. K. (1983). Topic shifts in couples' conversations. In R. T. Craig & K. Tracy (Eds.), *Conversational coherence: Form, structure and strategy* (pp. 136-156). Sage Publications.

Ditman, T., & Kuperberg, G. R. (2007). The time course of building discourse coherence in schizophrenia: An ERP investigation. *Psychophysiology, 44*(6), 991-1001. https://doi.org/10.1111/j.1469-8986.2007.00565.x

Ditman, T., & Kuperberg, G. R. (2010). Building coherence: A framework for exploring the breakdown of links across clause boundaries in schizophrenia. *Journal of Neurolinguistics, 23*(3), 254-269. https://doi.org/10.1016/j.jneuroling.2009.03.003

Dolón, R., & Sánchez, A. (1999). Agenda meetings: Topic progression in negotiating activity. *Estudios Ingleses de la Universidad Complutense, 7*, 49-67.

Garcia, L. J., & Joanette, Y. (1997). Analysis of conversational topic shifts: A multiple case study. *Brain and Language, 58*(1), 92-114. https://doi.org/10.1006/brln.1997.1871

Gardner, R. (1987). The identification and role of topic in spoken interaction. *Semiotica, 65*(1/2), 129-141. https://doi.org/10.1515/semi.1987.65.1-2.129

Gernsbacher, M. A., Tallent, K. A., & Bolliger, C. M. (1999). Disordered discourse in schizophrenia described by the structure building framework. *Discourse Studies, 1*(3), 355-372. https://doi.org/10.1177/1461445699001003004

Gogo, J. O. (2020). The contribution of education to moral decay in Kenya challenges and prospects. *International Journal of Education Humanities and Social Science, 3*(1), 20-32.

Graham, D. (2014). *KeyBNC* [computer software]. https://key-bnc.tfiaa.com/

Gray, B. (2010). On the use of demonstrative pronouns and determiners as cohesive devices: A focus on sentence-initial this/these in academic prose. *Journal of English for Academic Purposes, 9*(3), 167-183. https://doi.org/10.1016/j.jeap.2009.11.003

Hobbs, J. R. (1990). Topic drift. In B. Dorval (Ed.), *Conversational organization and its development* (pp. 3-22). Ablex.

Hoey, M. (1991). *Patterns of lexis in text*. Oxford University Press.

Iter, D., Yoon, J. H., & Jurafsky, D. (2018). Automatic detection of incoherent speech for diagnosing schizophrenia. *Proceedings of the fifth workshop on computational linguistics and clinical psychology: From keyboard to clinic* (pp. 136–146). Association for Computational Linguistics.

Jones, A. (2010). Why are logical connectives sometimes detrimental to coherence? In A. Mahboob & N. Knight (Eds.), *Appliable linguistics* (pp. 200-220). Continuum.

Just, S., Haegert, E., Kořánová, N., Bröcker, A., Nenchev, I., Funcke, J., Montag, C., & Stede, M. (2019). Coherence models in schizophrenia. *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology* (pp. 126-136). Minneapolis.

Kware, A. A. (2020). I will rather be poor but respectful than be rich but morally bankrupt. *International Journal of Arts and Social Science, 3*(2), 1-13.

Kyle, K., & Crossley, S. A. (2017). *Tool for the Automatic Analysis of Cohesion (TAACO): Quick start guide and user manual for TAACO 1.5*. https://www.linguisticanalysistools.org/uploads/1/3/9/3/13935189/_taaco_1.5_user_manual_7-16-17.pdf

Malekinejad, P. (2019). Fourth addenda to criticism of the current science in the world. *Scholars Journal of Arts, Humanities and Social Sciences, 7*(7), 474-481.

Malekinejad, P. (2020a). Sixth addenda to criticism of the current science in the world. *International Journal of Education Humanities and Social Science, 3*(1), 93-100.

Malekinejad, P. (2020b). Tenth addenda to criticism of the current science in the world. *EAS Journal of Humanities and Cultural Studies, 2*(5), 31-37. https://doi.org/10.36349/easjhcs.2020.v02i05.005

Meehan, T., & MacLachlan, M. (2010). Self construction in schizophrenia: A discourse analysis. *Psychology and Psychotherapy, 81*(2), 131-142. https://doi.org/10.1348/147608307X256777

Ndukauba, K. N. N. (2020). French language and national development the case of Nigeria. *The International Journal of Social Sciences and Humanities Invention, 7*(7), 6080- 6085.

Noël-Jorand, M. C., Reinert, M., Giudicelli, S., & Dassa, D. (1997). A new approach to discourse analysis in psychiatry, applied to a schizophrenic patient's speech. *Schizophrenia Research, 25*(3), 183-198. https://doi.org/10.1016/S0920-9964(97)00022-4

Panicheva, P., & Litvinova, T. (2019). Semantic coherence in schizophrenia in Russian written texts. *25th Conference of Open Innovations Association (FRUCT)*, 241-249. Helsinki. https://doi.org/10.23919/FRUCT48121.2019.8981535

Phillips, M. (1989). *The lexical structure of text: Discourse analysis monographs No. 12*. University of Birmingham.

Pojanapunya, P., & Watson Todd, R. (2018). Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory, 14*(1), 133-167.

Redeker, G. (1990). Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics, 14*(3), 367-381. https://doi.org/10.1016/0378-2166(90)90095-U

Riou, M. (2015). The grammar of topic transition in American English conversation. Topic transition design and management in typical and atypical conversations (schizophrenia). [PhD thesis, Université Sorbonne Paris Cité].

Rochester, S., & Martin, J. R. (1979). *Crazy talk: A study of the discourse of schizophrenic speakers*. Plenum.

Rogalski, Y., Altmann, L. J. P., Plummer-D'Amato, P., Behrman, A. L., & Marsiske, M. (2010). Discourse coherence and cognition after stroke: A dual task study. *Journal of Communication Disorders, 43*(3), 212-224. https://doi.org/10.1016/j.jcomdis.2010.02.001

Saavadra, J. (2010). Quantitative criteria of narrative coherence and complexity in persons with paranoid schizophrenia. *The Journal of Nervous and Mental Disease, 198*(5), 349-355. https://doi.org/10.1097/NMD.0b013e3181da47f5

SAGE and Alexander Street Press (2008). *Counseling and psychotherapy transcripts, client narratives, and reference works: A resource for mental health educators and practitioners*. Alexander Street Press.

Sanfilippo, A. (1998). Ranking text units according to textual saliency, connectivity and topic aptness. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics COLING-ACL, 98* (pp. 1157-1163). Montreal.

Scott, M., & Tribble, C. (2006). *Textual patterns: Keywords and corpus analysis in language education*. John Benjamins.

Stede, M. (2012). *Discourse processing*. Morgan & Claypool Publishers.

Watson Todd, R. (2003). *Topics in classroom discourse*. [PhD thesis, University of Liverpool].

Watson Todd, R. (2016). *Discourse topics*. John Benjamins.

Yule, G. (1996). *The study of language* (2nd ed.). Cambridge University Press.